User Behavior Analysis and Prediction Methods for Large-scale Video-on-demand System

Huimin Zhang
Abstract

User Behavior Analysis and Prediction Methods for Large-scale Video-on-demand System

Huimin Zhang

Video-on-demand (VOD) systems are some of the best-known examples of 'next-generation' Internet applications. With their growing popularity, huge amount of video content imposes a heavy burden on Internet traffic which, in turns, influences the user experience of the systems. Predicting and pre-fetching relevant content before user requests is one of the popular methods used to reduce the start-up delay. In this paper, a typical VOD system is characterized and user’s watching behavior is analyzed. Based on the characterization, two pre-fetching approaches based on user behavior are investigated. One is to prediction relevant content based on access history. The other is prediction based on user-clustering. The results clearly indicate the value of pre-fetching approaches for VOD systems and lead to the discussions on future work for further improvement.
## Contents

Part I: Introduction ....................................................................................................................... 7

1 Background and Scope .............................................................................................................. 8
   1.1 Background ....................................................................................................................... 8
   1.2 Scope and Limitations ..................................................................................................... 8

2 Research Questions .................................................................................................................. 9

Part II: Related Work .................................................................................................................. 10

3 Previous work ..........................................................................................................................11
   3.1 Characterizing user behavior on VOD system ................................................................. 11
       3.1.1 Access pattern ......................................................................................................... 11
       3.1.2 User interest distribution ....................................................................................... 11
   3.2 Pre-fetching scheme ........................................................................................................ 11
   3.3 Clustering based prediction ............................................................................................ 12
       3.3.1 Memory-based CF ................................................................................................. 12
       3.3.2 Model-based CF ..................................................................................................... 13
       3.3.3 Clustering methods ............................................................................................... 13
       3.3.4 Principal Component Analysis ............................................................................. 14

4 Data Description ....................................................................................................................... 15
   4.1 IPTV network .................................................................................................................. 15
   4.2 Definitions ....................................................................................................................... 15

5 Methods .................................................................................................................................... 18
   5.1 Statistical analysis .......................................................................................................... 18
   5.2 Primary prediction .......................................................................................................... 18
   5.3 Secondary prediction ...................................................................................................... 19
       5.3.1 Deciding K .............................................................................................................. 19
       5.3.2 Popularity-based prediction .................................................................................. 20
       5.3.3 Similarity-based prediction ................................................................................... 20
   5.4 Evaluation of prediction methods ................................................................................... 20

Part III: Sub-studies .....................................................................................................................21

6 Study 1: Quantitative study toward user watching behavior .................................................... 22
   6.1 Description ....................................................................................................................... 22
   6.2 Methods ........................................................................................................................... 22
   6.3 Results and Analysis ....................................................................................................... 22
       6.3.1 User Access Overtime ............................................................................................ 22
       6.3.2 User and Request Distributions ............................................................................. 24
       6.3.3 Probability of user’s next view event .................................................................... 25
       6.3.4 Potential for Prediction .......................................................................................... 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.5</td>
<td>Survey</td>
<td>27</td>
</tr>
<tr>
<td>6.4</td>
<td>Conclusion and Discussion</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>Study 2: User behavior study and prediction methods on channel T</td>
<td>30</td>
</tr>
<tr>
<td>7.1</td>
<td>Purpose and Data Description</td>
<td>30</td>
</tr>
<tr>
<td>7.2</td>
<td>Methods</td>
<td>30</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Primary Prediction</td>
<td>30</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Secondary Prediction</td>
<td>31</td>
</tr>
<tr>
<td>7.3</td>
<td>Results and Analysis</td>
<td>32</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Primary Prediction</td>
<td>32</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Secondary Prediction</td>
<td>33</td>
</tr>
<tr>
<td>7.4</td>
<td>Conclusion and Discussions</td>
<td>36</td>
</tr>
<tr>
<td>8</td>
<td>Study 3: User behavior study and prediction methods on channel M</td>
<td>38</td>
</tr>
<tr>
<td>8.1</td>
<td>Purpose and Data Description</td>
<td>38</td>
</tr>
<tr>
<td>8.2</td>
<td>Methods</td>
<td>38</td>
</tr>
<tr>
<td>8.3</td>
<td>Results and Analysis</td>
<td>39</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Potential of combining popularity and similarity based prediction</td>
<td>40</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Number of user clusters</td>
<td>40</td>
</tr>
<tr>
<td>8.3.3</td>
<td>Pre-fetching N programs</td>
<td>41</td>
</tr>
<tr>
<td>8.4</td>
<td>Conclusion and Discussion</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Part IV: General Analysis, Discussions and Acknowledge</td>
<td>44</td>
</tr>
<tr>
<td>9</td>
<td>Discussion and conclusion</td>
<td>45</td>
</tr>
<tr>
<td>9.1</td>
<td>General discussion and future Work</td>
<td>45</td>
</tr>
<tr>
<td>9.1.1</td>
<td>How could users be characterized in a large-scale Video-on-demand system?</td>
<td>45</td>
</tr>
<tr>
<td>9.1.2</td>
<td>To what extent can prediction methods be effective in improving user experience?</td>
<td>45</td>
</tr>
<tr>
<td>9.2</td>
<td>Conclusions</td>
<td>47</td>
</tr>
<tr>
<td>10</td>
<td>Acknowledge</td>
<td>49</td>
</tr>
<tr>
<td>References</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>
Part I:
Introduction

In this part, the background of the thesis project is introduced. What is video-on-demand system and pre-fetching system are explained in brief. Afterwards, the purpose of this study is explained and the two relevant research questions are raised.
1. Background and Scope

1.1 Background

This thesis project is conducted at Swedish ICT Acreo located in Kista, Stockholm, Sweden. It is under the project NOTTS, which aims at developing solutions and products for distribution of streaming media over Internet, so called over-the-top (OTT) distribution(1).

Video-on-demand (VOD) system is now a major step in the evolution of media content delivery, and it keeps gaining polarities in recent years because it provides users with easy access to the content(2). To meet its growing demand and consumer population, cable and satellite providers have already been in the process of developing on-line streaming video systems(3). Besides, cable VOD system also increasingly operates like on-line video service does, and increasing number of TV series and movies are becoming available through on-demand system for those cable subscribers with their set-top box.

With the growing popularity of VOD system, the importance of user experience becomes more and more significant. The huge amount of media traffic has imposed a heavy burden on the Internet. As a result, user sometimes needs to wait for the loading of the content before he could actually start to watch. Long access delay will obviously have a negative influence on the user experience. Relevant study reveals that the more familiar users are with the system, the more sensitives they are to the delays (4). A recent approach to reduce this user perceived latency is pre-fetching systems which proactively pre-load certain content to the cache before it has been requested by the user.

Most of the pre-fetching systems are based on user access history and content popularity. The first step of this prediction scheme is to characterize and analyze the user behavior using data mining techniques. Making a quality prediction has a direct impact on the performance of pre-fetching. Similar prediction methods are also used in recommendation systems where the recommended result needs to fit a certain user’s interest. After the prediction is made for user’s future watching behavior, the predicted contents are pre-loaded in the cache.

1.2 Scope and Limitations

This thesis project started at January 2015 and ends at June 2015 at Swedish ICT Acreo. It starts with the literature study and then mainly focuses on the characterization of an accessed large-scale VOD system and design the prediction methods based on user behavior.

There exists some limitations for this study. The characterization and prediction methods for the deployed large-scale VOD system is within the scope of the thesis project. However, small-scale or web-based VOD system is not involved so the conclusion of the study could not be representative for all the VOD systems. Besides, the cost of prediction is not in the scope of the thesis project.
2. Research Questions

In this thesis project, focus is on finding user behavior patterns of a large-scale VOD system to improve the user experience through reducing the start-up latency perceived by users. So the research questions are formulated as follows:

• How could users be characterized in a large-scale Video-on-demand system?
• To what extent can prediction methods be effective in improving user experience?
Part II:
Related Work

Recently, several studies have been carried out on one of the most popular Swedish TV providers who provide their on-line broadcasting content for their subscribed users. Some on-demand programs consist of a series of episodes. In one of that studies (5), a pre-fetching scheme is proposed in order to reduce the start-up delay. The prediction scheme predict which episode will be watched after the user has watched certain episode and pre-load the predicted contents into terminal cache. The result shows that the more contents to be pre-fetched, the higher potential in reducing the start-up latency. What also could be seen is the potential of pre-fetching system, especially in customized pre-fetching scheme for different demand patterns and video categories.

In this thesis project, the similar pre-fetching scheme is conducted for validation. Besides, clustering-based prediction is designed and conducted in order to verify its potential in VOD system. And data mining techniques are used to find behavior patterns and also to conduct clustering approach.
3. Previous work

3.1 Characterizing user behavior on VOD system

3.1.1 Access pattern

Understanding how this VOD system is accessed is usually the first and basic step for characterizing this system. This access pattern could be user access pattern, traffic pattern or the user’s active ratio, which is defined as the number of active hosts per minute in the analysis in (6). The study (7) is conducted on a large deployed VOD system in China. It starts with the user access patterns over time and the result from daily access pattern shows the peak in noon-break and after work. The similar two peaks pattern during the week is also found in another study (8). The user activity results in relevant study (6) and traffic patterns in (9) reveal a similar peak hour from approximately 7pm to 12pm. Besides, the daily user access patterns for different genres of programs are compared in (8) where the free content is the most popular throughout the day.

The result for weekly access pattern in study (7) shows a work habit during the weeks and frequent peak hours in weekends or holidays when users are relaxed at home. The average number of daily requests increases steadily in the second half of the week and reaches peak on Sunday. Unlikely, Friday and Saturday consistently had the lowest evening peaks within the week in the results from another study (8). These studies reveal that different VOD systems have different access pattern but share some part of characters.

3.1.2 User interest distribution

Pareto Principle is one of the most popular rule used to describe user interest distributions. The traditional Pareto Principle, also known as 80/20 rules, indicates that roughly 80% of the effects come from 20% of the cause (10). One of the example could be in business that 80% of the sales come from 20% of the clients. Related to VOD system, it means that 80% of the consumption data comes from 20% of the users. In study (7), a moderated 80/20 rule is revealed from the results of user request distribution which are spread more widely than predicted by the traditional Pareto Principle.

3.2 Pre-fetching scheme

Pre-fetching has became a popular technique to reduce the user perceived latency. The purpose of pre-fetching is to pre-load certain content to the cache before it is requested by the user. Among the current existing pre-fetching approaches, access-history based pre-fetching and popularity-based pre-fetching are widely used.

There are two steps in a pre-fetching system. The first step is prediction, which is about making prediction on the user’s watching behavior. After the prediction is
made, the pre-fetching engine will proactively store the video into a local cache. In this thesis work, focus is on the prediction methods so that the current network architects for involved VOD system will not be explained in details. In addition, assumption can be made that the pre-fetching is conducted at user’s end, which means that the pre-fetched content is stored in a terminal cache at user’s end. Similar assumption has been made in (5).

In study (5), a pre-fetching scheme is designed for a VOD system where the intrinsic structure of TV series is used and N nearby videos are pre-fetched to terminal devices. User’s history access patterns are collected to decide the nearby videos. The performance is evaluated by the terminal cache hit ratio and the results proves the potential of conducting pre-fetching and reveals that pre-fetching two relevant contents can reach 69% of the requested videos and has a optimal cost.

3.3 Clustering based prediction

The precision of the prediction is highly valued in a recommending system. To make a quality and personalized prediction, collaborative filtering (CF) has proved to be one of the most effective approaches, which uses the known preferences of a group of users to make recommendations or predictions of the unknown preferences for other users. In (11), a survey about different CF techniques is conducted, where the CF techniques are divided into three groups, memory-based CF, model-based CF and hybrid recommenders. In this thesis, only memory-based and model-based CF are involved and discussed.

3.3.1 Memory-based CF

Memory-based CF uses an entire or a sample of use-item database for prediction. Each user is a part of group who share the similar interests and the prediction of preferences on a new item for a user is based on his defined neighbors. A typical CF algorithm proceeds in three steps (12):

1. Calculating the similarity between two active users.
2. Neighborhood formation: select k similar users.
3. Generating top N items by weighted average of all the ratings of users in the neighborhood.

Neighbourhood-based CF is a prevalent memory-based CF algorithm. Usually the similarity between two users or two items are measured as the first step in this approach. And afterwards, the prediction is usually made by taking the weighted average of all the ratings.

Measurement of user similarity is the vital part who decides the accuracy of this collaborative filtering approach. There are quite a lot methods to measure the similarity or dis-similarity between two users.

Distances between two objects reveals the dissimilarity between them and Euclidean distance is widely used as one of the methods. The formula $d(P, Q) = [(x - a)^2 + (y - b)^2 + (z - c)^2]^{1/2}$ tells how Euclidean distance is calculated for two points $P(x, y, z)$ and $Q(a, b, c)$ in a three-dimensional space, which can also be extended to multi-dimensional space. Pearson correlation is a correlation-based
similarity measurement which measures the extent to which two variables linearly relate with each other. In comparison with Euclidean distance, Pearson correlation fixes the problem of grade inflation(13). When a user tends to give a generally higher ratings compared to another user, the result of their Euclidean distance will indicate that they are not similar while the results from Pearson correlation will still prefer their similarity.

Another approach to measure similarity is Cosine similarity. This method is designed for the sparse data like documents but can also be used for users and items. Two objects are treated as vector \( x \) and vector \( y \), and the cosine of the angle formed by these two vectors are computed through the formula 3.1.

\[
\cos(x, y) = \frac{x \cdot y}{\| x \| \| y \|}
\]  

3.3.2 Model-based CF

Models such as machine learning and data mining algorithms can allow the system to learn complex patterns by training data. Model-based CF is to make prediction based on learned models.

Clustering CF is the representative of the model-based CF techniques. Users are clustered based on their similar interests in purpose of making a quality prediction. Usually a cluster is a collection of objects that are similar within this cluster and are dis-similar to the objects in other clusters. The main advantage of this model-based CF is that it better addresses the sparsity problem of a data set. Meanwhile, it will lose useful information when the dimensionality are reduced(11).

In most situations, clustering acts as an intermediate step in the prediction methods and is used to partition the huge raw data. In this case, it is usually used together with other CF techniques like memory-based CF (14). The study (15) shows a significant improvement by a clustering based neighborhood prediction compares to the basic CF approach.

Not only users can be clustered, items can also be clustered into different classes based on their user groups. There are also some approaches, like (16), which combines user-clustering and item-clustering. In this study, users are first clustered based on their ratings on items and nearest neighbour of target user can be found based on the similarity. After this proposed approach, the item clustering CF is utilized, which leads to a more accurate result. Others approach like (14) clustered the users and items separately first, and repeated on re-clustering users and items where each user is assigned to a class with a degree of membership proportional to the similarity between the user and the mean of the class.

3.3.3 Clustering methods

There are some popular and simple techniques used as clustering methods, including K-means, Hierarchical Clustering and DBSCAN (17). DBSCAN is a density-based clustering algorithm that produces a partitional clustering, in which the number of clusters is automatically determined by the algorithm. Hierarchical clustering refers to a collection of closely related clustering techniques that pro-
duce a hierarchical clustering by starting with each point as singleton cluster and then repeatedly merging the two closest clusters until a single, all-encompassing cluster remains. K-means clustering is a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters (K), which are represented by their centroids.

K-means and Hierarchical algorithm are conducted on the same data set in a practice in (13), and from the result can be seen that K-means runs faster than the Hierarchical Clustering. In this study, K-means is chosen for its simple and fast-processing. The basic K-means algorithm can be described as follows(17):

1. Select K points as initial centroids
2. From K clusters by assigning each point to its closest centroid.
3. Recompute the centroid of each cluster
4. Repeat 2 and 3 until centroids do not change any more.

The initialization of centroids is random, which means that different runs of K-means will typically lead to different results. The limitation of random initialization can lead to the occurrence of empty cluster. When assigning a point to the closest centroid, which is in Step 2, Euclidean distance is often used for data points in Euclidean space, while cosine similarity is more appropriate for document type data.

(18) has purposed a possible augmented K-means clustering method which initializes the centroids with a simple, randomized seeding technique and obtain an algorithm that is O(log k)-competitive with the optimal clustering. Experiment results show that this augmentation improves both the speed and the accuracy of K-means.

3.3.4 Principal Component Analysis

Data set can have a large number of features, for example the total number of item available for the users is huge. In this case, dimensionality reduction is beneficial to reduce the total cost of computing and improve the clustering method. The reduction of dimensionality by selecting new attributes that are the subset of the old is known as feature subset selection or feature selection (17). Principal Component Analysis (PCA) is one of the dimensionality reduction techniques based on linear algebra approaches.

The main purposes of a principal component analysis are the analysis of data to identify patterns and finding patterns to reduce the dimensions of the data set with minimal loss of information(19). When deciding the optimal dimensions of subspace, explained variance works as a measurement. The cumulative distribution of explained variance can tell how many dimensions are enough to cover the major part of the features for current entire data space.
4. Data Description

4.1 IPTV network

In this thesis work, study is based on the data set from a Portuguese nation-wide IPTV network which is comprised of one month of consumption data. This VOD system provides user with a catch-up TV service which provides an access window of 7 days to approximately 80 TV stations, depending on user’s own subscriptions. It consists of 17 millions of access records from over 57 thousands of system users. The log data ranges from June 1\textsuperscript{st} to June 30\textsuperscript{th} of 2014. Each record includes both user’s information and detail information of the requested content. However, user’s private information like name, gender, birthday and so on are not accessible and included in this project. Pausing will not cause a new record, while stopping and resuming or restarting will cause a new record in the system.

Table 4.1 shows one sample entry record which covers most of the information listed in the data set. There are also some other information which is not listed in the sample record like 'IsHD', UTC information and 'ID' for each entry record, and those are not used in this thesis study. Since videos in this system have various types, so in this thesis project, videos are treated into two different groups, one is videos who belong to a certain program, like TV series, variety shows, news and so on, another is videos who do not belong to a specific program, like movies. Besides, some of the videos are re-broadcasted for several times during this recorded month.

The 'Account' is unique for each subscribers, while it is not the identifier for single device, which means each account can refer to several devices. Device type is described as 'IsPC' in each record. 'City' and 'District' indicate the geographical information of the account, and 'PlayTimeHour' is the specific time for user access. Information of the requested video content includes the broadcasting time of the content ('OriginalTime' and 'StartTime'), which channel this content belongs to ('ChannelCallLetter' and 'StationID'), the category information ('CachedThemeCode'), the title of the content and the series and episode information ('SeasonNumber' and 'EpisodeNumber'). The length of the video content can be calculated by the 'StartTime' and 'EndTime'. Besides, each video content has an unique identifier which is 'EpgPID', and related series video content belongs to can be identified with 'EpgSeriesID' within this channel.

4.2 Definitions

There are some definitions which need to be clear throughout this report.

**Video**(Episode) In this data set as mentioned, video is the actual broad-casted production. It could be an episode within a series, or an individual movie. In this data set, videos can be distinguished from each other by their unique 'EpgPID'.
### Program(Series)

As described in (20), program is a segment of content intended for broadcast on television. It could also be called as TV shows and it contains a series of videos/episodes. In this data set, programs can be distinguished from each other by 'EpgSeriesID'. The index within a program is shown as 'EpisodeNumber'.

### Request and Program request

In this data set, each access entry data can be defined as a request. Program request is the set of access entry data generated from a user for the same program. If a user has watched the same program several times or the restart of the system happens, then it will be defined as several requests but one single program request.

### Active User

In order to define active user, the user’s request distributions are analyzed beforehand. The result in Figure 4.1 shows that a modified 80/20 rule suits for the current data set. In this IPTV network, 30% of the user contribute to 80% of the consumption data, which is the red line in the figure. These top 30% of the users contribute to 75% of the program requests. The dashed red line in the figure indicates the user’s program request distribution, where the similar 80/30 rules applies. Then the active user is defined as the top 30% of the user who contribute to up to 80% of the program request. And those active user has covered around 77% of the requests.

<table>
<thead>
<tr>
<th>Account</th>
<th>DB94FD81D67FD10D1E07F24F790E1181</th>
</tr>
</thead>
<tbody>
<tr>
<td>PlayTimeHour</td>
<td>2014-06-07 08:43:52.287000000</td>
</tr>
<tr>
<td>OriginalTime</td>
<td>2014-06-06 19:30:00</td>
</tr>
<tr>
<td>City</td>
<td>Braga</td>
</tr>
<tr>
<td>District</td>
<td>Braga</td>
</tr>
<tr>
<td>Title</td>
<td>The Ultimate Spider-Man T2 - Ep. 11</td>
</tr>
<tr>
<td>IsPC</td>
<td>False</td>
</tr>
<tr>
<td>SeasonNumber</td>
<td>2</td>
</tr>
<tr>
<td>EpisodeNumber</td>
<td>11</td>
</tr>
<tr>
<td>EpgSeriesID</td>
<td>28849</td>
</tr>
<tr>
<td>EpgPID</td>
<td>635827</td>
</tr>
<tr>
<td>StartTime</td>
<td>2014-06-06 19:30:00</td>
</tr>
<tr>
<td>EndTime</td>
<td>2014-06-06 19:50:00</td>
</tr>
<tr>
<td>ChannelCallLetter</td>
<td>SICK</td>
</tr>
<tr>
<td>StationID</td>
<td>327260</td>
</tr>
<tr>
<td>CachedThemeTime</td>
<td>MSEPGC-Others</td>
</tr>
</tbody>
</table>

*Table 4.1. Sample Access Data*
Figure 4.1. Distribution of user’s request and program request
5. Methods

This project has been divided into 3 separate studies. Each study will use the current data set to some extent. Firstly, the statistical analysis is conducted on the entire consumption data. Afterwards, a more focused study is conducted on a smaller data set, which is the consumption data from a TV channel in this network called Channel T. Prediction methods are designed and used to pre-fetch relevant contents into cache in order to decrease the cache load and improve the user experience. Similar study will be repeated again on another TV channel called Channel M, who has a different character comparing to the channel T.

For Study 2 and Study 3, the prediction methods will be divided into two part, which is primary prediction and secondary prediction. As defined previously, video is the smallest unit and program is a series of videos. The result of primary prediction is relevant videos who belongs to a certain program, which answers which videos in this program will be watched after user has watched one episode in this program. In comparison, secondary prediction is for programs. The prediction results will be programs, which answers what relevant programs user will watch after he or she watched a program.

5.1 Statistical analysis

The project starts with a statistical analysis for the current VOD system. The purpose of this study is to know the basic watching behavior of the current user group and the information of current VOD system. Afterwards, a survey will be conducted on another user group to see the differences and validate the current findings in user behavior. As mentioned, the statistical analysis serves as the first sub-study in this project and detail methods and discussions will be explained in the next session - Study 1.

5.2 Primary prediction

As mentioned above, primary prediction is prediction method for videos who belongs to a certain program, like TV series, documentaries, variety shows and so on. In the current data set, videos within the same program share the same ‘EpgSeriesID’ and they are distinguished from each other by a unique ‘EpgPID’ and also by their episode number. So in the primary prediction, the prediction mainly focuses on those videos who have a ‘EpgSeriesID’ and also an episode number to indicate the index of the current video.

Primary prediction mainly answered which episodes the user will watch after he or she has watched a certain episode in a program. User’s previous access history will be investigated in order to see their behavior pattern. Previous work (5) has designed a pre-fetching scheme for this type of prediction. So in this project, this scheme will be validated.
It is assumed that user has watched an episode from a program, and studies about users next view event is conducted in order to get the prediction. Which episode in this program has the higher probability to be watched by this user needs to be figured out. After this study, a list of episode indexes will be generated and ranked by their probabilities. Afterwards, what also needs to be investigated is how many episodes with higher probability should be pre-fetched. The detail of primary prediction will be discussed later on in Study 2.

5.3 Secondary prediction

Secondary prediction is for predicting new content for users, which means that besides what user has already watched, what else this user may probably watch is the question to be answered by this prediction. Users are first clustered into different groups according to their watching behavior. The users who have a similar watching behavior will be clustered together and different clusters have different watching behavior to keep them distinguishable. In order to get users watching behavior, their access history will be collected. The clustering methods used in this study is augmented K-means (18) and will be mentioned as K-means in the later sections. Since the data set is huge and quite sparse, principal component analysis (PCA) could be used to see if it is possible to decrease the current dimension of the data and also help deciding the optimal number of clusters before conducting K-means clustering. After the users are clustered into certain groups. The similarity-based and popularity-based prediction is conducted within each cluster.

5.3.1 Deciding K

Since deciding K is a must for conducting K-means clustering. In this study, there are three methods which is conducted for deciding the optimal number (K).

The first approach is conducted on the raw data before clustering, which is the Principal Component Analysis (PCA). As described before in previous work, the purpose of PCA is to reduce the current dimension of the data space. The explained variance during PCA can provide with an suggestion on the optimal dimension for sub-data. How many dimension is needed can indirectly suggest about how many clusters could be optimal.

The second approach is about the clustering performance measurement. The distortion is one of the measurement, which is the sum of squared differences between the user's data point to centriod. The lower the distortion is, the higher the quality of clustering is. There are also other clustering performance measurements used to validate the results, for example Silhouette Coefficient, where a higher Silhouette Coefficient score relates to a model with better defined clusters.

The last approach is the most direct measurement for this study, which is the cache hit ratio performance of this clustering results. This directly tells how this K performs in this approach, however, the control of variance is difficult in this case, which means that several possible optimal K should be chosen through other measurements.
5.3.2 Popularity-based prediction

For popularity-based prediction, the weight for each program is the same for all the users who are in the same clusters. Because the popularity refers to how many percentages of users in this cluster has watched this program. Popularity could also explained as the probability of a user to watch this program. For program \( p \), the probability of being watched by user \( j \) is calculated by Equation 5.2 in this study. \( n \) is the total number of users in this cluster. \( w_{pi} \) indicates whether user \( i \) in this cluster has watched program \( p \). \( w_{j} \) indicates whether current user \( j \) has watched this program or not. If he has watched this program, then \( W_{pj} \) will be 0.

\[
W_{pj} = (1 - w_{j}) \cdot \frac{\sum_{i=1}^{n} w_{pi}}{n}
\]  
(5.1)

5.3.3 Similarity-based prediction

For similarity-based prediction, the similarity between each user is calculated by cosine similarity. And for each user, the weight of this program will be calculated as in the Equation 5.1 in this study. For user \( j \), \( W_{pj} \) is the weight of a program \( p \) to a this user. \( w_{j} \) implies if \( j \) has watched this program \( p \) or not. If user \( u \) has already watched this program, then the weight of this program \( p \) will have 0 as its weight because this is not new to \( j \). \( n \) is the number of the rest users in current cluster. \( w_{pi} \) indicates if user \( i \) in this cluster has watched program \( p \) or not. \( s_{ij} \) shows the similarity between user \( i \) and user \( j \).

\[
W_{pj} = (1 - w_{j}) \cdot \frac{\sum_{i=1}^{n} (w_{pi} \cdot s_{ij})}{\sum_{i=1}^{n} w_{pj}}
\]  
(5.2)

After the similarity and popularity prediction, for a specific user in his or her cluster, a list of programs ranked by their weights (\( W_{p} \)) will be generated.

5.4 Evaluation of prediction methods

After the prediction is made within each cluster, each user will have a lists of programs ranked by their weights.

For evaluating the performance of the prediction methods, cache hit ratio (\( H \)) is calculated. It is defined as the number of requests of videos (\( h_{r} \)) which are retrieved from the pre-fetching cache over the total number of requests (\( t_{r} \)). Higher hit ratio indicates that more pre-fetched contents are requested by users, thus indicating the better prediction performance.

\[
H = \frac{h_{r}}{t_{r}}
\]  
(5.3)
Part III:
Sub-studies
6. Study 1: Quantitative study toward user watching behavior

6.1 Description

In this study, the complete consumption data set are used, which covers 28 days of users access history over this nationwide IPTV network. The purpose of this study is to characterize the user’s watching behavior and find out some particular behavior for VOD system user. Whether the users are sharing interests among programs will be answered and thus proving the potential of prediction methods. Also getting familiar with the current dataset will be beneficial before conducting the prediction work on this user group.

6.2 Methods

In this study, the work will be divided into two parts. The first part is the statistic analysis towards the entire network. The work involves user’s access pattern overtime and the distribution of the access is characterized as a function of time, from across hours of the day to across days of the month. The current data set consists of 3 objects, which is users, channels and video content. The relationship between these object will be analyzed. And all of the studies will be analyzed in the distribution results.

In order to see if there is benefit to make prediction, whether this user groups share similar interests should be answered. One of the approaches is assuming the current user group share a proxy cache and once there is a user who has just watched a video, this video will be cached in the cache. If other user requested the same content, then the content will be directly loaded from the cache. In this case, cache hit ratio can also be the measurement to see the performance of proxy cache and the potential of prediction could be seen from the result.

After the statistic analysis, several findings will be gathered. In order to validate the users behavior and also preparing for the prediction, a survey is designed. The content of the survey is focused in the following part:

- The frequency of usage of the VOD system
- User’s access pattern over weeks
- User distributions among channels and programs
- Probability of user’s next view event within a program
- Zapping or surfing behavior

6.3 Results and Analysis

6.3.1 User Access Overtime

The study is started with the basic understanding of user access patterns in the system. The distribution of the access is characterized as a function of time, from across hours of the day to across days of the month.
Figure 6.1. Monthly pattern

Figure 6.2. Different daily access pattern for weekdays and weekends

Monthly Access Pattern

Figure 6.1 shows a big picture of the access pattern over the whole month, the data is aggregated every day and covers for 28 days in the month. There is no clear weekly pattern can be seen from the result, which may due to the National holidays and World Cup which were both taken place in June 2014. But the number of users is quite stable for the whole month.

Daily Access Pattern

Figure 6.2 shows the daily pattern of access, where a difference from weekdays to weekends can be seen from the result. During the weekdays, the system reaches peak hour in the evening, while in the weekends, the system seems to be active from the morning already.

Different Weekly Access Pattern

Additionally, different TV stations seem to have different patterns, which means that some of the stations have significant weekly patterns. For examples, Figure
6.3 shows the access pattern of a typical popular kid channel, where a big contrast between weekdays and weekends could be seen from the result. During the weekdays, there are two peaks in a day, the first one comes in the morning while the second and higher peak comes in the evening. In comparison, Figure 6.4 shows the weekly access pattern for another type of channel where series and variety shows are the majority of the content. For this channel, weekdays evening has a higher request number and more users comparing to the weekends.

6.3.2 User and Request Distributions

Overall, the requests each user generate differs from each other and also the popularity of each video differs. Pareto Principle, also known as 80/20 rules, is a popular rule used to describe the user interest distribution. To test if this system follows this rule, requests data is collected for each user and each program.

Users distribution

As can been seen from the Figure 6.5, 80% of the user has less than 50 requests for the entire month, which means less than 2 requests per day. The distribution of requests covers the whole data set. Some of the extreme active users generated more than 3 thousands video session requests over the month, but in most cases, user still tends to be not active in this data set.

As mentioned when defining active user, the distribution in figure 4.4 indicates that this user group fit a moderate 80/20 rule, which means that in this network,
80% of the consumption data is not contributed by 20% of user, but by 30% of the user instead.

Similar results are got from users distribution among videos and channels. 80% of the users watched less than 40 programs in a month. And 80% of the user watched less than 10 channels in the entire month. Around 20% of the user watched only 1 channel. What’s interesting to see is that the user’s number of request does not grow with the number of channels he watched. For example, users who has watched the 40 channels do not create more requests compared to the users who has watched only 10 channels. This somehow indicates that active users are more focusing on their favorite channels and those users who has watched a lot of channels is just less sticking to their previous behavior.

Request distribution over videos

Another analysis is conducted within different videos, where videos are grouped into ranks from 1 to 10000 by total number of views. The data covers the whole month and cumulative distribution result is shown in Figure 6.6. Here the result looks similar to a moderate 80/20 rule, or 90/10 rule, where 80% of the total video views comes to 12% of the total videos. This result clearly indicates that user’s interested is quite focused and this group of users share the same interest.

6.3.3 Probability of user’s next view event

As mentioned in the methods part, within each programs, prediction method is focused on predicting which episodes user will watch next and then pre-fetching
them when this video is broad-casted. Knowing what users next view goes is the starting point of this prediction method.

Result of this analysis on the entire network is shows in Figure 6.7, where 0 indicates that user would not watch any episode in this program after the current episode, which covers 30% of probability. If X is the current episode index, then most possible next view goes for index X+1, followed by X+2, X-1, X+3, X+4, X-2. Those videos who are independent and not belong to a certain program are not included in the analysis.

6.3.4 Potential for Prediction

Figure 6.8 shows the case when assuming that the proxy cache is used in the system and without any prediction. In this case, all the accounts share the same cache and once a video is watched, it would be saved in the cache. The result indicates a high hit ratio and a small fluctuation of hit ratio during the day. The cache hit ration reaches 99.7% within a day, which reveals that all users are requesting similar contents.
In order to exclude one of the possibilities that user keeps seeing the same content, another hit ratio is calculated for terminal cache where each account has its own cache. In the results, the cache hit ratio only reaches 20%. The big contrast between terminal cache and proxy cache indicates that user shares interests in this data set. This also reveals great potential for conducting prediction methods in the current VOD system.

6.3.5 Survey

The results of the survey are analyzed and compared with the statistic results. The survey is designed and conducted on 29 participants. Participants are recruited through internet and 25 of them are VOD system users and the results from those valid participants is analyzed. The majority of the participants are from China and Sweden, but nationality is not taken as a factor to be analyzed in this study. The range of age in the user group starts from 18 and biggest group is from 18 to 35. Both female and male participants are included but with a bigger amount of male. Families who has kids are also included. As for the occupation of this user group, bigger part of them are student, other occupations could be pensioner and other type of regular work.

The frequency of usage of the VOD system

Among these participants of the survey, when asked about how frequent they use the VOD system, 80% of the participants use less than 3 hours per day. Each time they use, most of the user will only watch 2 to 5 videos. Comparing with the IPTV network user, this user group tends to be more active because the statistic result from the data set is that 80% of the user watched less than 40 videos in a month, which is 1 to 2 videos per day. And 60% of the user from IPTV user group used the system less than once per day, while this frequency of usage only applies to 15% of the participant in this survey.

User’s access pattern over weeks

The result from the statistic analysis for IPTV network shows that weekends does not have the higher peak compared to weekdays after work. For some channels, weekday evenings even have bigger amount of request compared to weekend. The same result is found through the survey. When being asked about when they usually used the VOD system, up to 80% of the participants in the survey chose weekday evenings, followed by weekend afternoons and weekend evenings. Bigger part of participants chose to use VOD system both in weekdays and weekend, while there is even one participant who only use the system in weekdays evening after work.

User distributions among channels and video

In the statistic analysis, 80% of the user only watch less than 10 channels and only around 4% of the total video contents have been watched by users. So in this survey, user’s interests in channels and videos are studied. Participants are asked about how many favorite channels they have and to what extent they will only focus on those channels. The result is that more than 80% of the participants have 1 to
4 favorite channels and over half of the participants tend to focus on their favorite channels when they are using the system. Compared to the statistic results, these participants seem to be more focused because they have fewer favorite channels while they are more active in using the system.

Similar questions are also asked about the programs instead of channels. Most of the participants in this survey have 1 to 5 favorite programs, the rest participants has bigger amount of favorite programs that he or she usually watch. This number is slightly more than the number of their favorite channels. When being asked about to what extent they will only focus on their favorite programs when they use the system, participants tend to have similar (little higher) loyalty to their favorite programs, comparing with results from channels.

**Probability of next view event within a program**

In statistic results, the highest probability of next view event goes to the next episode. In current IPTV network, there are also some users who requested the same content for several times, some of them are in continues viewing session, some others are not. Whether user actually has the behavior of watching the same content again is also involved in the survey. Since the content of the video may influence the users behavior, so that 2 typical type of programs are mentioned separately in the survey. The first one is the TV series, where each episode is displayed in sequence and they have a strong connection with the previous one and the next one. Another type of the programs is the variety show, where each episode is more independent from each other.

In the survey, when talking about TV series, 45% of the participants will not watch anymore unless they are interested in this series. And $\frac{2}{3}$ of the participants will probably watch the next episode. Only 8% of the participants will watch the same content again. In comparison, after having watched an episode from a variety show, $\frac{1}{3}$ of the participants will not watch anymore unless they are interested in this variety show. Less than half of the participants will watch the next episode in this variety show. Jumping between different episodes are more common when they are watching variety shows. What’s more, only 1 participant thinks he or she will watch the same content again, which indicates that repeat viewing same content is not a common behavior. The differences in user behavior between TV series and variety shows suggest the value of customized pre-fetching approach towards different content.

**Zapping or surfing behavior**

As mentioned in previous session, zapping or surfing behavior happens in VOD system. Whether it is an common behavior is studied through this survey also. What’s also included is what will cause a user’s watching behavior. However, zapping behavior is not included as one of the factor in the further studies.

Participants are asked how long they would probably stay when they switched to a program that they do no think it’s interesting. This question was raised in order to see if there could be a time limitation to define the zapping behavior in this VOD system. The result is that they will usually stop when they cannot stand it any longer, and usually it is within 5 minutes. This result is longer than the most of the program surfing time in previous studies, which indicates the potential of
longer viewing session being excluded while the user is actually surfing. Because length of viewing session is not reachable in this data set, so this factor is not taken into consideration in the later studies.

Participants are also asked what reason is for them to start watching a video. The recommendations from family or friends are the biggest reason for watching a video. Another big part of the reason is that they have been watching it for quite a while. Then the rather big part of the reason is that they get to know this program through different media so that they would like to have a try. Only a small part of participants will watch a program spontaneously. This result indicates that users are focusing on their favorite programs and the recommendations play an important role when they would like to watch something new.

6.4 Conclusion and Discussion

From Study 1, there are several conclusions according to the current IPTV network and the behavior of VOD system user. For this nationwide IPTV network, firstly, the difference of access pattern is not clear. However, the difference between weekdays and weekends is quite obvious, where weekends has a earlier arrival of peak hour. Weekdays evening even higher request number even than weekend, which has also been validated through the survey.

Besides, different channels have different characters in their daily patterns. Difference among channels is not only about daily access pattern, but also about their contents and the assignment of category of their video content. This difference implies the risk of conducting the same prediction approach. Taking all the channels individually and customize the prediction methods towards different channels will be optimal approach for this VOD system.

As for the user’s watching behavior, user’s cumulative distribution fit a moderate 80/20 rules, which is 80/30 rules. Video’s cumulative distribution also fit a moderate 80/20 rules, while which is 90/10 rules. These results show that focusing on those top 30% of the user can already cover up to 80% of the consumption data, and those active users are more valuable when conducting prediction. Another essential part in user’s watching behavior is their next view events, as the statistic result shows, the highest probability is to watch the next episode. Besides, there is quite high possibility that they will not watch any more in this program. Then the result from the survey indicates a slight different next view event because of different content of the program.

At last, the potential of conducting prediction has been proved by the proxy cache hit ratio result which implies that users in this IPTV network share the same interest.
7. Study 2: User behavior study and prediction methods on channel T

7.1 Purpose and Data Description

From Study 1, the conclusion could be made that different TV channels in this network have different characters and they are different from each other in the aspect of their contents, their user groups and how they assign their contents into certain groups. This difference brings the value to taking a specific channel out from this entire network and conducting prediction on this channel instead of the entire network. So the purpose of this study is to see how the prediction could work for a certain type of channel.

In this study, one of the featured channels is selected from the nationwide IPTV network. According to the statistic analysis from this IPTV network, channel T is the TV station who has the biggest amount of access logs and users, which is over 3 millions logs and over 270 thousands users. It covers around 18.7% of the total access logs in the complete data set.

The contents in channel T have a wide range of categories, including TV series, variety shows, kids programs, movies, news and so on. Among those categories, TV series is the most popular content for the users, which covers up to 90% of the total logs. As conclusion, this TV channel can be categorized as a TV series channel.

As for its user group, 80% of the users for this channel has less than 25 logs in the entire month, which means biggest part of the user watched this channel less than once per day. In order to focus on the most valuable user behavior, this study is focusing on the active users in this channel, which is over 80 thousands users. The definition of active user is covered in the previous session. For each active user, his or her 20-day’s complete data is collected.

7.2 Methods

Prediction is divided into two part in this study. As the majority of program contents in this channel are TV series. So that which episode in a program the user will probably watch next will be the primary prediction and what new programs the user will probably watch will be the secondary prediction. The performance for prediction is represented by the cache hit ratio. As mentioned in data description, 20 days of consumption data is collected for this study. First 10 days is used for making prediction and the next 10 day’s data is used to evaluate the performance of this prediction approach. In the first 10 days, there are altogether 44 programs. The prediction will be generated within the current range of programs and the evaluation of performance will be measured by cache hit ratio described previously.

7.2.1 Primary Prediction

For prediction within a TV series, which is the primary prediction in this study, user’s viewing behavior is collected and analyzed and then the prediction is made
based on their access history. In details, in the first 10 day’s consumption data, what user’s next view goes needs to be studied. Only the videos who belong to a certain program will be taken into account. How many relevant episodes in the same program should be pre-fetched will be answered by the result of cache hit ratio.

7.2.2 Secondary Prediction

Secondary prediction is mainly for prediction between programs, which aims at predicting what user will watch if they would like to watch new programs. Both user-based and item-based clustering methods are used as an intermediate step for this prediction. For this prediction, user’s watching pattern is collected from the first 10 day’s data. Because different programs have different number of episodes broad-casted during this time period, which means that user’s interests in this program cannot be simply evaluated by the number of episode watched or access generated for this program. So in the end, user’s watching pattern is defined as a series of boolean values for each program.

Table 7.1 shows an example of users watching pattern data. \( n \) is the total number of programs the cluster members have watched, which is between the value of 0 and total number of programs in this 10-days period (44). \( P_1 \) to \( P_n \) indicates programs. Each user has a watched or not watched behavior towards each program, which is represented by 1 and 0 respectively in the table content.

Firstly, since each user has a 44 dimension featured data as his or her behavior pattern. The purpose of item-clustering is to decrease the dimension of the current behavior data. So the PCA analysis is conducted to see how many dimensions are enough for this big data set. The results will also suggest how many clusters for programs are enough. After the items are clustered using K-means clustering method, each user will have a new watching pattern for clusters of programs instead of programs themselves. What this new pattern contains are not boolean values but percentages of viewing for programs in this program cluster instead. Then the user is clustered based on their new watching pattern also by K-means clustering methods. The number of clusters is decided by the distortions for different number of clusters.

Table 7.2 shows the new behavior pattern data after the video clustering is conducted. \( m \) is the total number of program clusters generated. \( C_1 \) to \( C_m \) indicates program clusters. Each user has a value towards each program cluster, which is

<table>
<thead>
<tr>
<th>Account</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
<th>( C_5 )</th>
<th>( C_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account_1</td>
<td>100%</td>
<td>30%</td>
<td>0%</td>
<td>20%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Account_2</td>
<td>100%</td>
<td>20%</td>
<td>0%</td>
<td>10%</td>
<td>0%</td>
<td>5%</td>
</tr>
<tr>
<td>Account_3</td>
<td>20%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>50%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 7.1. Sample behavior pattern data

Table 7.2. Sample new behavior pattern data after video clustering
represented by how many programs in this program cluster he or she has watched. Taking Account_1 as an example, he or she has value of 100% in C_1 column, which means that he or she has watched all the programs in Cluster 1.

After the users are clustered into a specific cluster, both popularity based and similarity-based prediction are conducted within each cluster. How many programs need to be pre-fetched to reach a good prediction performance will be answered through different cache hit ratio results. For secondary prediction, the prediction is based only on a program or a series, which means that the exact index of episode is not taken into consideration. Additionally, the program or series this user has already watched will not be in the prediction list because secondary is for predicting what the user will watch if he watches something new.

7.3 Results and Analysis

In this TV channel, primary prediction and secondary prediction have their individual results. The comparison between them could be implied by hit ratio since the total number of access request is the same for both of them.

7.3.1 Primary Prediction

The first step in primary prediction is to know what user is going to watch after he or she has watched an episode in a series. As mentioned previously, a series could be drama, variety shows, documentary and so on. And the previous 10 day’s data is used for this analysis. As can be seen from the results in Figure 7.1, if we assume that a user watched episode X, with over 60% of the probability that he will watch the episode with index X+1. 0 indicates that user will watch any episode in this program after he has watched episode X, which has around 13% of the probability. From this figure, we can see that the indexes of episode with higher possibility to be watched are: X+1,X+2,X-1,X+3,X+4.

Comparing the current result from channel T with the result from same probability study for entire IPTV network (Figure 6.7), the probability of watching next episode is much higher in channel T than the probability in entire IPTV network. There could be several reasons. Firstly, the user groups are different. The current user group is the active users who use the system more frequently than the average level. This active status might cause this higher result. In addition, channel T is the TV series channel. As mentioned in the survey results in Study 1, the viewers of a TV series have a higher possibility to watch the next episode than the viewers of a variety show. The content of channel T is then another reason for this difference.

Deciding on pre-fetching how many relevant episodes is the next question which needed to be figured out. The performance of pre-fetching from 0 to 5 relevant episodes is evaluated through terminal cache hit ratio and shown in Figure 7.2. Pre-fetching more than 5 relevant episodes is not taken into consideration because it will cost heavy load to the cache and the improvement is limited. Pre-fetching 0 episode means that only the terminal caching is conducted and only the watched episode from this account will be saved in the cache. As can be seen from the blue line in Figure 7.2, 21% hit ratio can be achieved in the first day when con-
ducting terminal caching only and without any pre-fetching. Combination with pre-fetching the next episode, whose index is X+1 will bring the improvement of 17% in hit ratio. This result proves the potential and performance of primary prediction. As can be seen from Figure 7.2, pre-fetching more than 3 relevant episodes will only bring slight improvement of cache hit ratio. This result indicates that pre-fetching 3 episodes, which are X+1, X+2, X-1, is optimal for current system and can reach 43% cache hit ratio.

Taking a look at the green line which indicates the cache hit ratio for the second day, a bigger improvement can be reached when combining terminal cache with pre-fetching the next episode, which is 24% of improvement in hit ratio. Compared to the results from the first day, the overall performance is increased by 10% in average, which can reach 52% hit ratio when pre-fetching 3 relevant episodes and combining the terminal cache. This results also implies the great potential of pre-fetching 3 relevant episodes.

What need to be mentioned is that when conducting terminal caching without any pre-fetching. The cache hit ratio reaches up to 8% in the beginning. This illustrates that there exists some continues short viewing session for the same content and these sessions are not caused by pausing the video. So these continues session for the same content within an hour will be preferably regarded as some system problem. This also reveals that The cache hit ratio for pre-fetching 0 is higher than normal. Additionally, the performance of 0 pre-fetching has a bigger fluctuation compared to the other results which means the hit ratio is fluctuated.

7.3.2 Secondary Prediction

After the behavior pattern of each active user is collected, each user has a 44 dimension featured data as their behavior pattern. K-means clustering are used for item clustering, which is video clustering in this case. Before clustering, PCA analysis is conducted on the current raw data. The results for explained variance is shown in Figure 7.3. 19 dimensions is good enough for the current 44 dimensions data because it has kept up to 95% features of the current data. Then 44 programs
are clustered into 19 clusters using K-means. And a new user behavior pattern is generated and each user has a 19 dimensions data.

On this new behavior pattern data, another PCA analysis is repeated to see how many user groups are good enough for user clustering. The result is that the number of dimensions has a linear increase as the growing explained variance, which suggests that the number of necessary dimensions is not clear enough using PCA analysis. Then the distortions for clustering results is used in order to decide how many clusters are good enough for this new pattern data. The green line in Figure 7.4 shows the distortions according to different number of clusters. The decrease becomes relatively small after 60, which suggests 60 as a optimal number for good quality clustering. The blue line in Figure 7.4 shows the distortion result if the video clustering is not conducted, which means that each user has his or her original 44 dimensions behavior data. Compared to the result in green line, the new data has a lower distortions with the same number of clusters, which means that fewer clusters can get the same quality of clusters if the video clustering is conducted before user clustering.

After the clustering is conducted. Both similarity and popularity based prediction is measured by using the following 10 day’s data. After trying different num-

---

**Figure 7.2.** Cache hit ratio of pre-fetching N episodes

**Figure 7.3.** Explained variance for the raw behavior data
ber of pre-fetched programs, pretty bad performance in terms of low hit ratio was got. And in the following 10 days, there were altogether 9 new programs broadcasted and those new programs cannot be predicted so that those access logs for new programs are excluded in this measurement. In Figure 7.5, the green line and green dots shows the performance of popularity-based and similarity-based prediction respectively when pre-fetching 10 new programs. What need to be mentioned is that pre-fetching 10 new programs will be quite a heavy load to the cache and the performance is still not good. The similarity-based prediction can only reach a 2% cache hit ratio in the beginning and popularity-based prediction can only reach 10% hit ratio. In comparison, the hit ratio in primary prediction is much higher (45%) even when pre-fetching only 1 new episode. And since the secondary prediction is only prediction for programs instead of the exact index in the program, which means that the performance will be even worse if the prediction is made for an exact episode.

After investigating the relevant factors which might cause the worse performance. Users behavior pattern and other relevant information for channel T is collected from the following 10-days data. There are several reasons which cause
the low hit ratio. One of the biggest and most significant factors for this performance is that users tend to stay within their previous watching patterns. This indicates that they prefer watching what they always watch and not starting to watch a new program very often. This behavior exists also in the user group for Study 1. When talking about the reasons to watch a program, most of the participants in the survey tends to focus on his favorite programs and channels when he would like to watch something. As can be seen from Figure 7.6, even for those active users from channel T who contributed up to 80% of the consumption data, there are 42% of the users who will not watch any new program. Over 90% of the active users watched less than 2 new programs, even including those newly broad-casted programs in the following 10 days. The reason for their loyalty to previous watched programs could be that channel T is basically a TV series channel.

When focusing on those users who actually watched new programs, which is 58% of the current active user group, the same prediction methods are measured. The red dashed line and red dot in Figure 7.5 shows the result. This result indicates that even when assuming that all the users will watch some new programs, the hit ratio for this prediction methods could only reach 16% for popularity-based prediction and 4% for similarity-based prediction, under the condition of pre-fetching 10 new programs.

7.4 Conclusion and Discussions

In this study, primary prediction and secondary prediction are conducted on the most popular TV station in this entire IPTV network. This channel is mainly broad-casting TV series. For primary prediction, users history access patterns are analyzed and the highest probability of their next view event goes to the next episode. Afterwards, how many relevant episodes should be pre-fetched is investigated through their performance in terms of cache hit ratio and the conclusion is that pre-fetching 3 relevant episodes can reach a optimal performance. For secondary prediction, 10 days of user’s watching behavior patterns are collected, and each user has a 44 dimensions data to character his or her behavior pattern in the first 10 days. In order to decrease the current dimensionality, the programs are
clustered into 19 clusters and user’s new behavior patterns are generated which has 19 dimensions instead. Afterwards, users are clustered into 60 clusters according to their new behavior patterns. After clustering, both similarity and popularity prediction are conducted within each cluster.

Comparing the prediction performance of primary prediction with that of secondary prediction, the conclusion could be that the primary prediction could work as the main prediction method for channel T because it reaches a better performance and users tend to stay within their previous watched programs. And in primary prediction, pre-fetching 3 relevant episodes is optimal and reaches good performance. However, the cost model for this pre-fetching is not investigated so that deciding on the optimal number of pre-fetched relevant should also be validated based this. Combining the secondary prediction will only bring slight improvement to the prediction performance and at the same time bringing extra load to the cache since much more contents will be pre-loaded. The results of primary prediction also validate the findings and methods in (5) and prove the great potential of combination of pre-fetching and terminal caching. However, the category of this program is not taken into consideration when conducting primary prediction, whether the actual content of this program will affect the performance of the same prediction method could be interesting to see.

For the negative results from the secondary prediction, there were several reasons which might cause this performance. It is difficult to predict newly broad-casted contents because the future play list is not accessible in this project. Dealing with these new programs, current method is to exclude them. Another reason mentioned in analysis is that users focuses on their previous watched programs. This phenomenon may be caused because the content of the programs in Channel T is TV series. It is not surprising that users are more loyal to TV series. There could also be reasons like the quality of clusters also. Whether this methods of clustering is qualified should be tested, and also this bad performance of prediction is caused by the content of the program should also be studied, which brings up the value of Study 3.

Except for the bad performance, there is something interesting about the result, which is the pattern of cache hit ratio. As can be seen from the Figure 7.5, it is not hard to understand that the first day has the highest hit ratio. But which cannot be neglected is that there is another relatively high hit ratio happens after 7 days from the first peak. As mentioned, the current video-on-demand system offers a catch-up TV service where the content will be available for 7 days after it has broad-casted. Whether this 7-day catch-up service has an impact on the performance is also a good approach to get the reason for the second peak.
8. Study 3: User behavior study and prediction methods on channel M

8.1 Purpose and Data Description

As discussed and analyzed in Study 2, primary prediction can work as the main prediction method when conducting prediction on channel T. In order to test the performance of the secondary prediction and the quality of clustering method, Study 3 is conducted on a movie channel where the primary prediction does not work. In this channel, only secondary prediction will be conducted.

The selected channel is channel M, which is the most active movie channel according to the number of requests and number of users. The contents in this channel is movies, but the type of the movies are variety. The same 20-day’s of consumption data for active users are collected. As previously defined, the number of active users are more than 26 thousands in this channel, which is 33% compared to the number of active user from channel T. However, there are close to 300 programs during this 20-day’s period, which is much more than those 44 programs in channel T. This difference is mainly because the movie is regarded both as an individual program and video while TV series is only a program consisting of several videos.

One of the biggest reasons behind the unsatisfying performance of secondary prediction in Study 2 is that users are sticking to their previous watching patterns and only half of them will watch something new. In order to make sure about whether this factor will also influence the clustering-based prediction in channel M, a preliminary study is conducted to compare the watching pattern for the first 10 days and the next 10 days. Figure 8.1 shows the cumulative distribution of the number of newly-watched programs in the next 10 days for channel M. Almost all the user will watch new programs in the next 10 days and 80% of the user watch 5 or less new programs. Compared to the result in Figure 7.6, a conclusion could be that users for channel M are less loyal to programs and they will not stick to their previous watching patterns. After excluding this factor, the performance of clusters can be validated more in this channel.

8.2 Methods

In this study, the prediction will be focus on the secondary prediction, which is the clustering-based prediction approach. The consumption data from the same period is collected, the first 10 day’s data is used for clustering and prediction and the next 10 day’s data is used to evaluate the performance of this prediction approach. In the first 10 days, there are more than 280 programs and the prediction is generated within this range of programs.

The sample watching behavior data is the same as shown in table 7.1. K-means clustering methods will be used for conducting clustering. The number of clusters can be decided by distortions or by PCA analysis. In this study, a differ-
ent approach will be made instead of distortions or PCA analysis, and also item-clustering will not be included. The reason behind the changes comes from the results of PCA analysis and distortions result.

The first modification is not including the item-clustering in this study. As mentioned in Study 2, the purpose of conducting item-clustering before user-clustering is to decrease the current dimension of data. However, the PCA analysis results show that it is not optimal to decrease the current dimension, because the explained variance has a linear increase with the growing number of dimensions. Another result in distortion also indicates the same suggestion, where the distortion has a similar linear decrease when the number of program cluster grows. For those two reasons, the item-clustering is not included in the current study.

Another modification in method occurs when deciding the number of user clusters. The result in distortions has a similar linear decrease as the distortions of program clusters. This linear decrease makes it a challenge to select an optimal number of clusters. In this case, the new approach is to try two different number of clusters first and then compare the performance. The selected numbers of clusters is 80 and 40, which is selected from the distortion results.

After users have been clustered into certain group. Both similarity and popularity based prediction will be conducted within each cluster. Besides, the combination of those two prediction methods will also be conducted. The combination prediction methods is the unit set of the popularity methods and similarity methods. The evaluation of performance will be measured by cache hit ratio described previously. And how many programs to pre-fetch is also determined by their cache hit ratio performance.

8.3 Results and Analysis

The first step in this prediction approach is to decide the number of user clusters. Afterwards, both similarity and popularity prediction methods are conducted within each cluster. The performance of these two methods are analyzed and also the potential of combining these two methods are also investigated.
8.3.1 Potential of combining popularity and similarity based prediction

In order to know if it is beneficial to combine popularity-based and similarity-based prediction, how much improvement this combination will bring to the cache hit ratio results is analyzed. In this case, the number of user clusters is assumed as 80 and number of pre-fetched program is 50. Pre-fetching that much programs will bring the load to the cache but will also make sure the optimal performance of prediction, which is the reason for choosing 50 as the starting point. As can be seen from figure 8.2, the green line on the top is the performance combining popularity-based and similarity-based prediction. The dashed line and the dots line refer to popularity-based and similarity-based respectively. The combination of two prediction methods only brings 1% improvement in terms of cache hit ratio compared to the performance of popularity prediction. This result reveals that conducting popularity prediction can be the optimal solution for this case and combining with similarity prediction is not beneficial in channel M.

What also needs to be mentioned is the different performance in popularity-base and similarity-based prediction. The results from Figure 8.2 implies that the popularity prediction has a much better performance comparing to similarity prediction, with highest hit ratio over 30%.

8.3.2 Number of user clusters

As mentioned in method description, number of user clusters in this study is determined by their performance instead of analysis in user’s watching behavior data. 80 and 40 are selected as the number of user clusters, with the support of the PCA analysis result and distortions for different number of user clusters. The same group of users are clustered into 80 and 40 groups respectively. The performance of these two number of clusters are illustrated in Figure 8.3. The orange part is the performance for 40 user clusters and the green part the for 80 user clusters. This figure reveals that they have similar performance, and the performance of 40 clusters are even slightly better than the performance of 80 clusters. Especially in similarity prediction, the performance of 40 clusters are always better than that of 80 clusters. This indicates that 40 clusters is a better choice for channel M.

![Figure 8.2. Cache hit ratio with different prediction methods](image)
8.3.3 Pre-fetching N programs

When deciding the optimal number of pre-fetched programs, the popularity-based and similarity-based prediction are taken separately for analysis. These two prediction methods seem to have different performance as the different performance pattern over the 10 days also.

**Popularity-based prediction**

Figure 8.4 is the performance evaluation for different number of pre-fetched programs with popularity-based prediction. The decrease in performance is close to even when the number of pre-fetched programs decreases by 10. The decrease occurs from 50 programs to 40 programs is slightly smaller than the rest, which indicates that pre-fetching 40 programs could be an optimal choice. However, 40 programs is still a big amount for the cache.

The overall performance pattern for popularity-based prediction over 10 days is decreasing. The highest hit ratio occurs in the first day and the performance keeps decreasing in the next 9 days. One of the reasons for this decrease can be that some of the programs in the previous 10 days are no longer accessible after a week and there are also some new programs which cannot be predicted in this approach.
Figure 8.5. Cache hit ratio for pre-fetching N program with similarity-based prediction

Similarity-based prediction

Figure 8.5 is the performance evaluation for different number of pre-fetched programs with similarity-based prediction. The different performance between 50 programs and 40 programs are much smaller than the rest, also the difference in performance between 40 programs and 30 programs are also relatively smaller than the rest. But since the overall performance for similarity-based prediction is worse than that of popularity-based prediction, so pre-fetching 30 programs for similarity-based prediction is not an optimal choice for channel M.

The performance pattern for similarity-based prediction over 10 days are different from that of popularity-based prediction. Even though the cache hit ratio is also decreasing in similarity-based prediction, it has the fluctuation in the later 9 days instead of keep dropping.

8.4 Conclusion and Discussion

In this study, only secondary prediction is conducted on the another TV channel in order to see the performance of clustering-based prediction approach. Channel M is a popular movie channel in the nationwide IPTV network. Primary prediction is not applicable for movie channels because there usually is not any episode in movies.

For secondary prediction, the approaches in this study are also different from Study 2. Item-clustering is not included and the number of user clusters is determined by their performance instead of results from PCA analysis. Users are clustered into 80 and 40 clusters separately using K-means clustering. A similar and even better performance are got from 40 clusters, which means that 40 clusters can be an optimal number for this study.

Not only similarity-based and popularity-based prediction, but also the combination of them is conducted and evaluated. The results are that popularity prediction has a much better performance in terms of cache hit ratio comparing to similarity prediction. And combining both of them only bring around 1% improvement in cache hit ratio in comparison of popularity-based prediction. These results indi-
cate that the popularity-based prediction makes the most of the contributions so that the prediction methods should focus on it. Overall, the highest performance as 30% is still not an ideal result for this study.

The performance pattern for popularity-prediction over the next 10 days is decreasing. The highest hit ratio occurs in the first day and the performance keeps decreasing abruptly in the next 9 days and reaches only 1% in the end. There are several reasons for this pattern. Firstly, some of the programs in the previous 10 days are no longer accessible for users. Besides, there are also some newly broadcasted programs. Dealing with these new programs, current method is to exclude them. But if this prediction method needs to be put into practice, broadcasting new content is quite normal and happens every day. This brings up the issue of updating the clusters and predictions to make sure that all the programs can be predicted. One of the possible solutions is conducting updates on clusters and predictions everyday so that the performance can keep at the highest level everyday.

Another thing which is interesting to think about is the different daily patterns for popularity-based prediction and similarity-based prediction. In comparison, the performance pattern for similarity-based prediction has a bigger fluctuation instead of dropping. This also happens in Study 2 and the reason for this is worth studying.
Part IV:
General Analysis, Discussions and Acknowledge
9. Discussion and conclusion

In this thesis work, the study has been divided into 3 sub-studies. Each of the studies has its own methods and they share some methods in common. Each of them has their own results also. After these 3 sub-studies, the research questions raised in the beginning can be answered and also there exist some future works raised from these studies.

9.1 General discussion and future Work

There are two research questions raised in the beginning of the study. Study 1 mainly answers the first research question and Study 2 and 3 answers the second research question.

9.1.1 How could users be characterized in a large-scale Video-on-demand system?

The purpose of Study 1 is to characterize the user behavior in a large-scale video-on-demand system. As can be seen from the weekly access pattern, the highest peak in the entire week happens at weekday’s evening instead of weekend’s. This is a different result compared to the similar study \(^7\), and the difference indicates a different character towards different user groups. For this current accessible VOD system users, they tend to be more active after work and there is not a second peak after noon in weekdays. And this user group has a different watching behavior towards different channels also.

For the result of user distributions towards requests, 30% of the user covers around 80% of the total consumptions, which is also a moderated 80-20 rules. And the difference indicates that users tend to be less active than they suppose to.

User’s next view event indicates user’s watching behavior, for the entire user group, around 36% possibility will go for the next index. While among the users in channel T, the possibility of watching next episode is 60%. The reason for this difference can be that channel T is a TV series channel, so that watching in sequence is more natural in this type of TV channel. While for the entire data set, there also exist some variety shows or documentaries where the index sequence is no longer relevant, which lower the possibility of watching next episode. This difference has also been proved by the result from the survey.

9.1.2 To what extent can prediction methods be effective in improving user experience?

As can be seen from the results of Study 2 and Study 3, two different channels have significantly difference in their characters. Channel T is the TV series channel and has a large user group but smaller number of programs, while channel M is a movie channel and has a relatively smaller user group but more programs. Then
the pre-fetching approaches towards those two channels get different results also. For channel T, where both primary prediction and secondary prediction are applicable, the big difference in prediction performance can be seen. Primary prediction performs better than secondary prediction. However, when conducting prediction on channel M, where the primary prediction is no longer applicable, the performance of secondary prediction proves the value of the clustering approach. These differences between two channels reveal that the potential of clustering-based prediction depends highly on the different content of programs or characters of those TV channels. The result from quantitative study also reveals the different watching behavior toward different content.

In the future, how the different pre-fetching approach can be made for different program contents can be studied. In the current accessible resource, the content of the programs can not be easily characterized within a same standard. This limitation excludes the program categories as one of the factors to be analyzed. However, the character of a TV channel is relatively clearer, which supports the division of study.

As for the current clustering-based prediction method, there are several ways which can improve the performance.

Firstly, improving the quality of raw data is beneficial to achieve the optimization. Currently, only 10 days consumption data are collected to analyze the user’s watching behavior. Longer collection period would bring the better and more accurate description of user’s watching behavior. Besides, there is a limitation about the accessible consumption data, which is the missing information about the length of a viewing session. This limitation indicates that a very short viewing session can not be easily seen from the data set. As mentioned previously, zapping behavior is not within the scope of this study. While taking the zapping or surfing behavior into account will influence the clustering results also. Excluding the short viewing session will decrease the noise in current data set.

Another factor for the current raw data is the boolean value for each program. Most of the recommender systems use ratings from a user to illustrate his degrees of interests. While in VOD system, the degrees of interests cannot be simply represented by the number of requests generated for this program. Some of the program may be daily broad-casted and some are weekly broad-casted, which makes the request number as measurement not justifying enough. So finding the correct relation between degrees of interests with requested number is an issue for the future study. Besides improving the quality of raw data matrix, involving the item clustering actively can also optimize the clustering approach. Due to the time limitation, the total variance in Study 2 is controlled so that the item-clustering is not involved. Combination and repeating re-clustering has been proved to bring improvement in related studies so that this can be one of the future approach for clustering-based prediction.

Secondly, the clustering approach is mainly used for recommendation system. In the current study, only prediction is involved which means that users are not aware about those prediction results. Studies have proved the positive impact of recommendation upon user behavior (21). So that combining this clustering-based prediction with a recommender will highly increase the probability of the predicted
program being watched by users because users tend to watch the recommended items.

Last but not least, the current clustering approach is basically local and offline clustering. The shortage of this approach can be seen from the patterns of performance, where the terminal cache hit ratio drops down after the first day and reaches only 1% in the 10th day. In order to keep the performance in the highest level always, updating is important. One of the approaches can be updating the clusters and prediction results everyday. Besides, the popularity-based prediction has a higher performance, which means that the popularity of this program can be an important factor towards this clustering approach. If the prediction can be made about whether a certain program is going to be popular among the cluster members, then the quality of prediction can also be improved.

9.2 Conclusions

With the growing population of video-on-demand system, the extra load for the internet traffic is a significant issue. Some of the users will meet start-up delay because the requested content is still being loaded. Reducing the waiting time for users will definitely improve the user experience of current existing VOD system. Pre-fetching techniques are popular methods used to solve this issue. The scheme for pre-fetching is to predict users watching behavior and then pre-load the content to the local cache, thus reducing the waiting time for users.

In this thesis work, focus is on making quality prediction for a large-scale VOD system with the purpose of improving the user experience. A Portuguese IPTV network are involved in the study. In order to make a valid prediction on the users behavior for this IPTV network, characterization of user’s watching behavior is analyzed as the first sub-study in this thesis work. The first significant character is the different watching behavior for weekdays and weekends. Comparing to weekends, more users are watching at weekday’s evening. Users also have different weekly watching patterns towards different TV channels. Also in this study, a survey is conducted on VOD system users in order to validate those characters and preparing for designing the prediction methods for this VOD system.

After this study, a pre-fetching approach is conducted on the most active channel in this IPTV network. Predictions are divided into two part, primary prediction and secondary prediction. Primary prediction is to predict which episodes in this program are going to be watched after a user has watched certain episode in this program. Secondary prediction answers what programs user will watch if he watch something new. The result is that primary prediction has a much higher performance comparing to secondary prediction. There are several factors which may contribute to the result. TV series as the majority of content from this channel is one of those factors, which leads to user’s loyalty to their previous watching habit. Only half of the users will watch new programs, another half will only watch the newly broad-casted episode from the programs they had watched. So the conclusion can be made that the prediction methods for a TV series channel should be focused on primary prediction, which can already get a good performance in terms of cache hit ratio.
In order to see the performance of clustering-based prediction, another channel is selected where the users will watch new programs because most of the programs are movies. In this channel, the primary prediction are no longer applicable so that the performance of secondary prediction can be measured. The results show the potential of conducting clustering-based prediction on this type of channel, where the hit ratio can be around 40%. Also two different prediction methods after clustering is conducted, which is similarity-based prediction and popularity-based prediction. In comparison, the popularity-based prediction has a much better performance in terms of cache hit ratio. The combination of those two methods are also conducted where only 1% improvement can be got in the results so that it is not cost-worthy to make this combination.

Overall, the conclusion could be made that the prediction methods should be dynamic enough for different channels or different program content. Different prediction approaches can be conducted towards different user watching behavior also.
10. Acknowledge

This thesis would not be complete without the useful help from my supervisor Christina Lagerstedt and reviewer Mats Lind. And it is also greatly appreciated that my colleagues from Swedish ICT Acreo provided me with their great support and suggestions, especially Manxing Du, who also supervised me and helped me a lot in this project.
References


